Image Captioning Error recognition using Deep Attention based models

Abstract

In this paper we use modern Deep Learning approaches to solve a problem that requires fine-grained interaction of vision and language inputs. We use an image captioning dataset having incorrect captions with one incorrect word. The two tasks we handle are caption classification as right or wrong, and given an image with incorrect caption find the incorrect word. Existing results on the dataset are using non-Transformer based approaches. We use two types of models to solve both tasks. First, classification models and second, image captioning models. We create multiple baselines and try different architectures using RoBERTa, BERT, GPT2 for both types of models and verify their usefulness in solving the required tasks. We also access the usefulness of vision encoder-decoder models for image captioning. For this challenging dataset, we are able to achieve SOTA results for the classification task but not for foil word detection task. We analyse all approaches, the tradeoffs we made, and show example caption generations. Our work indicates that off-the-shelf general models are ideal for creating vision and language classification models but image captioning needs more fine-grained, dataset specific modifications which may not be generalizable. We release our code for further exploration and research in this area.

1. Introduction

Computer Vision (CV) and Natural Language Processing (NLP) fields have historically developed individually but now there are increasing applications like multi-modal speech assistants and voice-controlled robots where both need to be used in a collaborative manner. The tasks involving intersection of both fields are referred to as 'AI complete' tasks because they represent ideal information exchange between the two most popular modalities. Two types of tasks in this category are Visual Question Answering (VQA) and Image Captioning (IC). VQA tasks generally have an image and an associated question. The answer could be given by looking at the image and is usually short or has a few sentences. Other varieties of answers include fill in the blanks and multi-choice questions. We could additionally have access to a Knowledge Base (KB) which assists in giving answer about the image not explicitly in the image itself. IC problems deals with writing multiple correct captions for a given image. The captions should be correct, informative, and highlight the most important aspects of the image. With the popularity of these tasks, multiple datasets have been proposed in each domain. For VQA we have [1], [14], [15], [6] and for IC we have [2], [8], [21].

Figure 1. Sample image in dataset.
True Caption: A baby laying in a crib with a stuffed teddy bear
Foil Caption: A baby laying in a crib with a stuffed teddy cat
Note: The foil word here is cat, and the model needs to identify this from the image.

There has been great success in creating models that can effectively tackle these datasets but contrary to what one would believe, it has lead to the uncovering of another problem - these datasets are not challenging enough to test a vision and language models. Many VQA datasets can be answered without even looking at the image or by trivially concatenating image and language features. For IC too, models cannot distinguish between a correct an incorrect caption. Humans can do such tasks easily and we would want an AI complete system to also do it easily and efficiently. Given this, the problem of effective integration of vision and language inputs is far from solved. [17] proposed a new challenging dataset FOIL-COCO which is a modification of MS-COCO [12] having incorrect captions

Akshay Sasikumar, Amit Gopal Hattimare

for some images. The captions have been created in a way to make it difficult for the models to trivially identify the error. They introduce three problems. The two which we address in this paper are caption classification (T1) and incorrect (foil) word detection (T2). In T1, we have an image and one of the many captions for that image. Our task is to identify if the caption correctly represents the image or not. Incorrect captions have been created by replacing one word with an incorrect word. The replacement is done for important nouns, objects, items that are central to the image. Incorrect words have been chosen from the set of items in the dataset belonging to the same distribution. The intention is to make the task very difficult to solve by just reading the caption. In T2, given an image and incorrect caption, we need to identify the incorrect word. The third task, which we do not tackle but mention for completeness is to find the correct replacement of the foil word. Effective models should be able to integrate image and language inputs, locate the caption objects in the image, identify the incorrect word, and determine the correct word from a portion of the image. They show what many then SOTA models perform poorly on this task. We approach T1 and T2 with recent advancements in both NLP and CV domain and access how they can overcome the issues mentioned in [18].

We create two types of models in our experimentation - classification and image captioning and then use each of them to solve both T1 and T2. For classification, we first start with a simple baseline approach using ResNET image embedding and BERT caption embedding followed by fully-connected (FC) layers and softmax layer. We try multiple variations here by freezing the encoder models and only fine-tuning the linear layers, training end-to-end where the base models are also fine-tuned, adding different activation functions, trying different learning rates, using batch normalization, using regularization like dropout, among some other hyperparameter tuning. Our second approach is to use vision encoder, text decoder model followed by classification layer. We use Vision Transformer based encoders and BERT/GPT2 based decoders as they are SOTA in their respective domains. For image captioning, we do not pick an existing SOTA IC model but create one using vision encoder and text decoder models. Even here we use multiple encoder-decoder architectures and compare their performances. The details of these models along with how they are used for T1 and T2 are explained in section 3.

2. Related Work

With the advancement of vision and NLP models, there is an increasing research interest in 'AI complete' tasks like Visual Question Answering (VQA)[24] and Image Captioning (IC).

In image captioning, the task is to automatically generate a caption for an image, which is fluent and semantically correct with relation to an image. The idea is that the model learns about the image and is able to produce an informative caption that is relevant to the image. A popular dataset used for this is the MS COCO dataset [12]. A wide variety of approaches have been used to solve this problem, including Visual Commonsense Region-based Convolutional Neural Network [22] which involves a novel unsupervised learning method where causal intervention is used to allow a model to learn sense-making knowledge which is extra from the patterns in the data and pre-training using image text pairs for learning cross-modal representations in Oscar [11]. The latter particularly relates to the work done here in using combining pre-trained image and text embeddings together on the paired datasets.

VQA is a semantic task that aims to answer questions based on an image. It aims to understand and improve image understanding in generating answers to questions. The belief is that open ended diverse question answering problem is a good metric to verify whether a model is learning information available in both images and the question provided. The VQA dataset is fairly popular and has had a large number of approaches being applied to it like [25], [9], [20], [19] to list a few.

The FOIL-COCO [18] dataset is an extension of the MS-COCO [12] dataset which has been built with the purpose of ensuring that models that work well in VQA are actually understanding what is present in an image, and not just using the question itself to provide likely answers. It creates captions which are changed by just a single word and whether a model can identify it as the wrong word. Shekhar et al. [18] shows that VQA models of that time do not perform very well on this task. We aim to apply modern deep and pre-trained networks on this task to understand how well they have learned the meanings and semantics of the image and text they have been trained on.

Deep networks have been shown to work well on other tasks such as Image classification, such as ResNet [7]. It has shown that deeper networks, though more difficult to train, is able to store a lot of information in its representations and is able to perform well on ImageNet [3]. We look to seeing how well these representations have learned information about the image and apply it to the Foil task.

In a similar vein, pre-trained representations of text from deep networks have been a popular approach in recent years. Transfer learning has seen huge success in NLP fields recently, in models such as BERT [4] and T5 [16]. T5 has shown that pre-training on a large enough corpus of data, can give a huge flexibility to tasks that can be done using just prompts provided to the model. We look to finding out if the knowledge learned by such large models, can be applied to the FOIL tasks by encoding images in a fashion that can be used by large pre-trained NLP models.

An approach to such a model has been done in [13],

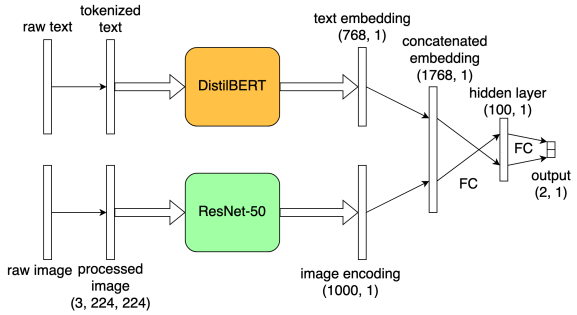

Figure 2. Architecture using BERT and ResNet to get image embeddings and then concatenating them before passing through FC layer and softmax.

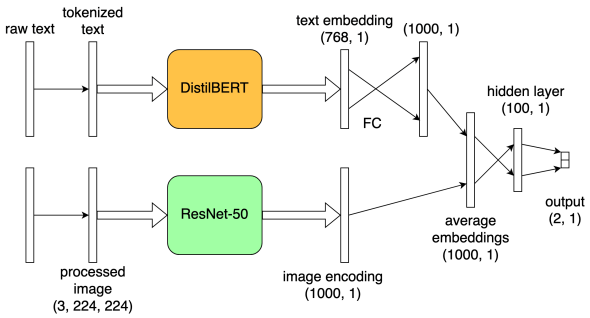

Figure 3. Architecture using BERT and ResNet to get image embeddings and then averaging them before passing through FC layer and softmax.

which looks to create a pretrained model with groundings between vision and language. This which will give learnt representations for transfer learning, rather than having this grounding be done at train time specific to a task. This model consists of co-attentional transfomer layers on both image and text input and could have representations which can be used for FOIL tasks, but this has been left as future work for now.

3. Approach

We mention details about all the model architectures tried in this paper. Then we detail how we use these models for tacking T1 and T2 tasks.

3.1. Model Architectures

3.1.1 Embedding based Classification Models

Baseline ResNet + BERT

This is a composition of DistilBert and Resnet-50 pre-trained models followed by fully-connected layers as seen in figure 2. For creating the text embeddings, we use the pre-trained DistilBertTokenizer with truncation and padding on. As all of our captions are small (∼50 words),

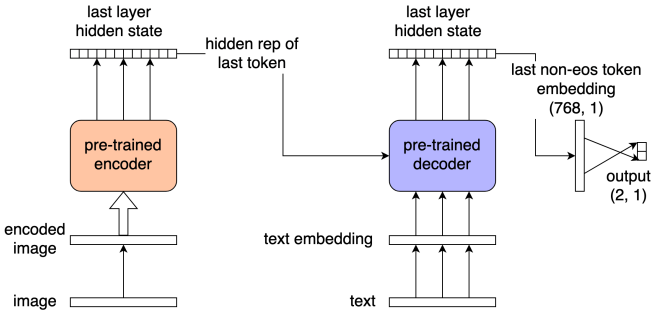

Figure 4. Architecture using Encoder to get image embeddings and Decoder to take a caption for classification.

we don't end up truncating any caption and use 0 padding to convert every caption to a fixed sized tensor. We pass the tensor to the BERT-based model and get embeddings for all the input tokens. We only use embedding of the `[CLS]` token and discard other token embedding following standard practice. For image, we first pre-process the image by resizing to `(3, 224, 224)` pixels, centering and normalizing as expected by the ResNet-50 model. Then we pass it through the model and get `(1000, 1)` sized logits tensor because ResNet is trained on ImageNet data that has 1000 output classes.

Two embedding architectures of this model were tried.

1. We use both these embeddings and concatenate them to create a tensor of shape `(1764, 1)`.

2. We convert the BERT embeddings from `(764, 1)` to `1000, 1)` using a feed forward layer. Now with both these embedding in the same shape of `1000, 1)`, we average them to create a final embedding tensor of shape `(1000, 1)` as well. Seen in 3

These embeddings are finally passed though a fully-connected layer to get intermediate output of shape `(100, 1)`. After a non-linearity, it is again passed it through an FC layer to get `(2, 1)` logits output which is then finally passed through a softmax function to get probabilities for both classes. We use cross-entropy loss as is the standard in many classification problems.

3.1.2 Encoder-Decoder Model

The effectiveness of pre-trained Transformer based models to be used for encoding both images and text is shown in [10]. We use Hugging Face library's [23] VisionEncoderDecoderModel class to create a custom foil classification and image captioning (image to text sequence generation) models using existing pre-trained image encoders and text decoders. We pre-process the image by resizing to `(3, 224, 224)` pixels, centering and normalizing. Then pass

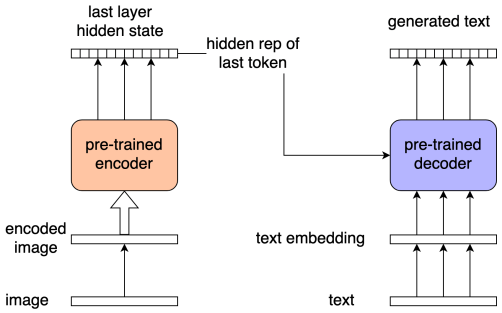

Figure 5. Architecture using Encoder to get image embeddings and Decoder generate an Image caption (for T2).

it through ViTFeatureExtractor [5] to get image encodings. We try `google/vit-base-patch16-224-in21k`, `google/vit-base-patch16-224` pre-trained ViT models. Both are pre-trained on ImageNet-21k (14 million images, 21,843 classes) at resolution 224x224 but the latter is further fine-tuned on ImageNet 2012 (1 million images, 1,000 classes) at resolution 224x224. For getting text encodings, we use tokenizer specific to the decoder architecture. We try RoBERTa, BERT and GPT2 pre-trained tokenizers for text encoding and corresponding decoders for text generation. We set the max length for tokenizer to 20 seeing our dataset captions.

Encoder-Decoder Classification Model

We use the above Encoder-Decoder model to generate representations of the captions. The input image is encoded by the encoder and passed as the first hidden state to the decoder. The decoder takes in the caption as an input, and we take the representation of the last word of the caption to build a liner classification head on top of. For each caption, we apply the linear classification layer and create scores for two classes. Using a standard classification cross entropy loss, we train the encoder-decoder model to perform classification on whether a given caption for an image is foil or not.

Encoder-Decoder Image Captioning Model

For Image Captioning, the HuggingFace VisionEncoderDecoderModel forward pass gives a single loss value and logits scores for each generated word over the vocabulary V. We use this to perform gradient update and use IC for T1 and T2.

3.2. Using Models for our tasks

3.2.1 Classification models for T1

Classification models can be directly used for T1. In all our models, we have a linear layer at the end giving 2 logit scores. We use a softmax layer on top of it to get 2 probability scores for the two classes - foil and non-foil. We use cross-entropy loss using the target labels for training the models.

3.2.2 Classification models for T2

We can use the trained binary foil classification models to figure out which word in the caption is the foil or not. So, we do this by masking out individual words in a given foil caption. For each word, we use the classification model to calculate the score for whether the masked caption is a foil caption or not. We calculate the score for all masked captions, and the masked caption with the lowest probability for being a foil caption will be related to the foil word. This is because the score for being a foil caption should drop by the most when the foil word is removed from the caption. This method is known as the occlusion method, which has been included in the methods applied in [17].

3.2.3 IC model for T1

To use IC for T1 we use approach similar to [17]. For a given image, we get captions from the IC model. Let the caption be $(w_1, w_2, w_3, ..., w_n)$. Then for each word w_i, we replace it by the word v_i most probable according to the probability distribution of the IC decoder language model. Then we get an updated caption as $(w_1, w_2, ..., v_i, ..., w_n)$. For all such n captions produced by replacing each of the n words one at a time, we get conditional probability using the LM and compare it with the conditional probability of the original input caption. If any of the n new captions has a higher probability than the original caption, then we mark the original caption as foil. We hypothesize that the new caption with updated word v_i is better for the image. We assume this because we know that the original data set has foil captions where only one word has been altered. The probability for any word is less because of large vocabulary size V (50k for GPT2 model). This makes the conditional probability of the caption a very small number. Do prevent numeric underflow, we use sum of log of token probabilities to get probability of a caption and use that for comparison.

3.2.4 IC model for T2

We perform this experiment only on foil captions i.e., we know that there definitely is one foil word in the caption. Given the original caption C, we find the best updated caption $\hat{C}$ where one word w_i has been replaced by v_i similar to the classification approach. If we cannot find any such $\hat{C}$ then we cannot find the word but if we do find a $\hat{C}$ then the word being replaced w_i is our identified foil word.

4. Experiments and Results

This section begins with what kind of experiments you're doing, what kind of dataset(s) you're using, and what is the

Actual Caption	Predicted Caption
"a room filled with wooden table and chairs"	"a group of birds on a bus with a person inside."
"little girl looking down at leaves with her truck with training wheels parked next to her."	"a group of birds on a bus with a person inside."
"a couple of young men sitting in front of a child's keyboard."	"a group of birds on a bus with a person inside."
"a kid using a spoon eating from a plate"	"a group of birds on a bus with a person inside."

Table 1. Some predictions from validation set using ViT encoder, Gpt2 decoder. The decode is not able to produce variable captions.

way you measure or evaluate your results. It then shows in details the results of your experiments. By details, we mean both quantitative evaluations (show numbers, figures, tables, etc) as well as qualitative results (show images, example results, etc).

4.1. T1: classification

For classification, we split the training data into train and validation set with ratio 9:1. We use validation set for hyperparameter tuning and for finding the best model. To check correctness of the model, we initially trained it with small data of size 20 and made sure that the train loss approached zero and accuracy approached 100%. We use AdamW optimizer with default parameters as it gives good convergence. We use various learning rates in the range of $[5e-4, 2e-3]$ at different times to speed up convergence given initial convergence speed. We pick the model giving highest accuracy on the validation data and report test data accuracy as the evaluation metric. The test results for different models are present in table 2. Along with overall data accuracy, we also report individual accuracy of foil and non-foil examples to compare performance in those cases.

The best result in [18] paper was obtained from HeiCoAtt model. It gave 64.14 accuracy but foil caption accuracy was only 36.38 showing the challenging nature of dataset. Our deep learning approach based on simply concatenating or averaging the image and text embeddings give performance comparable to HeiCoAtt model but still lower than that. We achieve high validation accuracy of 96.44 but that doesn't generalize to the test data. This could be because even though we use different train and validation splits, we split them randomly from the original train set so the images are shared across these two datasets. Finally, the encoder-decoder based classification model performs the best in our case. It's validation accuracy is comparable at 93.95 but the best improvement is seen in test data performance. The accuracy of 89.13 is much higher than all previous approaches. It also excels in foil data classification by achieving 88.84 accuracy compared to less than 40 for others.

4.2. T2: foil-word identification

The results can be seen in table 3 using both classification and image captioning models. Here, our approach does not beat the existing HeiCoAtt model. In fact, there is a huge gap in performance even though the same classification model (google-vit-base-patch16-224-in21k encoder and gpt2 decoder) performed very well in T1. We did not use embedding concatenation/averaging based classification model for this task as it did not perform well in T1. For IC, we tried a variety of encoder and decoder combinations but none worked as per expectations. The best IC model was obtained using google-vit-base-patch16-224-in21k encoder and gpt2 decoder. But even there, the quality of caption generation was not good. We analyse some examples in table 1.

We see that the IC model produces similar captions for various examples even if they are completely unrelated. The model is perhaps not able to generalize well and/or show variability in results. In the table all the captions (with their respective different images) produce the same output caption. We found that the produced caption exactly matches with one of the training captions which shows that the model is learning training data captions and not able to produce new sentences.

4.3. Data Analysis

We are using the FOIL-COCO dataset. The dataset itself is a structured dataset in JSON format which contains a list of image URLs with IDs and a list of captions associated with an image ID. We pre-processed this data to retrieve a flat csv dataset which contains image file name, image Flickr url, image dimensions, the caption and a boolean column about whether the caption contains a foil word, i.e., a modification of the caption or not. This boolean column is to be used as the label for T1.

The structured JSON data can be downloaded here. Some stats about the dataset are:

- Contains links to 97,847 unique images.
- Contains 297,268 captions and 297,268 corresponding foil captions.

Model Architecture	train examples	val overall	val foil	val non-foil	test overall	test foil	test non-foil
HieCoAtt	200,000	-	-	-	64.14	36.38	**91.89**
DistilBERT+ResNet+Concat+Linear	8000	96.44	94.75	98.16	60.58	37.13	84.02
DistilBERT+ResNet+Avg+Linear	8000	83.26	-	-	-	-	-
Encoder-Decoder+Linear	15815	93.95	97.09	89.87	**89.13**	**88.84**	89.42

Table 2. Test and validation accuracy for different architectures along with breakup of foil and non-foil caption accuracy.

Model type	Encoder	Decoder	Train Examples	Test Accuracy	Type
Classification	HieCoAtt	-	200,000	**33.69**	-
Classification	google-vit-base-patch16-224-in21k	gpt2	15815	7.92	best val
IC	google-vit-base-patch16-224-in21k	gpt2	785	8.68	last epoch
IC	google-vit-base-patch16-224-in21k	gpt2	785	10.13	best val
IC	google-vit-base-patch16-224	gpt2	785	9.28	best val
IC	google-vit-base-patch16-224	roberta-base	785	8.32	best val

Table 3. Foil word detection accuracy on test data (829 examples) for different image captioning architectures.

- The images are not in standardized sizes and vary for each example. There are some gray scale and 4 channel images too which we prune out.

- The mean width of the image is 577 pixels with the mean height being 485 pixels. The min and max for each dimension are: Width(59, 640), Height(51, 640). They are loaded and pre-processed to a standard size of `(3 x 224 x 224)` for loading into ResNet model.

- Train and test splits are included.

- Train data includes 65,697 image urls, and 197,788 caption and corresponding foil pairs.

- Test data contains 32,150 image urls, and 99,480 caption and corresponding foil pairs. Test data also includes the foil word for the foil caption, i.e. which word was changed. This will be used for checking T2 accuracy as well. This foil pair data is not available at train time.

  Due to resources constraints, we have limited to using a subset of the dataset for all of our training and testing, as reported in 3

For data processing, images are downloaded once, prior to training, based on the image urls available in the dataset.

4.4. Compute Resources

We used CPU resources from Macbook Pro 2.6 GHz 6-Core Intel Core i7 for small data training and development. For training transformer based models we use GPU resources from Google Colab and Amazon EC2 p2.xlarge instance having one Tesla K80 GPU with 12GB memory.

5. Conclusion

Modern models have come a long way since the FOIL COCO dataset was released, and larger models pre-trained on very large datasets are able to encode the information to be able to perform well on what is considered a relatively hard task. Using an encoder-decoder model with large pre-trained image encoder and text decoder models has shown to work well and produce good classification models on the foil task.

In the future, we would like to extend this work to be able to better solve T2 as well, i.e identifying the foil word, and further figuring out what the actual word that was replaced was. This would show that the model has learnt about the full information that is in the image, and left out in a foil caption.

Further, recent work has been done on building jointly pre-trained Language and Image models [13] whose learnt representations could have even more potential in a such a problem with closely related images and captions being used.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1

[2] Xinlei Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 4

[6] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. *arXiv preprint arXiv:1505.05612*, 2015. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[8] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 1

[9] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021. 2

[10] Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models, 2021. 3

[11] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2

[13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 2, 6

[14] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27:1682–1690, 2014. 1

[15] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2015. 1

[16] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019. 2

[17] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*, 2017. 1, 4

[18] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurelie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. "foil it! find one mismatch between image and language caption". In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages 255–265, 2017. 2, 5

[19] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 2

[20] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[21] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1

[22] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020. 2

[23] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 3

[24] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017. 2

[25] Deshraj Yadav, Rishabh Jain, Harsh Agrawal, Prithvijit Chattopadhyay, Taranjeet Singh, Akash Jain, Shiv Baran Singh, Stefan Lee, and Dhruv Batra. Evalai: Towards better evaluation systems for ai agents. *arXiv preprint arXiv:1902.03570*, 2019. 2